Semantic Segmentation of Built-up Areas in Satellite Imagery: A Deep Learning Approach with Comparative Analysis

C. D. Pace, A. Bria, C. Marrocco, M. Molinara, V. Rega
Dept. of Electrical and Information Engineering
University of Cassino and Southern Lazio
Cassino, Italy
European University of Technology EUt+
European Union
{cesaredavide.pace; a.bria; c.marrocco; m.molinara; vincenzo.rega}@unicas.it

M. Focareta, G. Meoli
Mapsat srl
Benevento, Italy
{m.focareta; g.meoli}@mapsat.com

***Abstract*—The increasing anthropisation of natural environments, particularly in urban contexts, necessitates accurate and scalable tools for territorial monitoring to support smart city development. This work presents an automated semantic segmentation framework based on deep learning for identifying built-up areas in high-resolution satellite imagery. The proposed system employs a UNet architecture and explores multiple encoder backbones, including VGG, ResNet, EfficientNet, and the Mix Vision Transformer (MiT) family. A comprehensive evaluation reveals that the MiT-B2 encoder achieves the highest performance, with an F1-score of 0.862 and an Intersection over Union (IoU) of 0.757. Experimental results highlight the critical influence of ground truth annotation quality on model accuracy. Compared to traditional semi-automatic methods, the proposed approach significantly reduces false positives and enhances the delineation of complex anthropized structures, such as road networks. These findings demonstrate the potential of advanced deep learning solutions for operational satellite-based monitoring of urban expansion and land use, supporting data-driven decision-making in smart city governance.**

***Index Terms*—Built-up Areas, Semantic Segmentation, Deep Learning, UNet, Satellite Imagery, Anthropisation Monitoring**

I. Introduction

The transformation of natural environments by human activity, known as anthropisation, has led to a significant increase in built-up areas such as cities, industrial zones, and infrastructure. This expansion poses several environmental challenges, including habitat loss, pollution, soil sealing, deforestation, and contributions to climate change. Consequently, monitoring these anthropised regions is crucial for sustainable land management. Satellite imagery offers a valuable tool for this purpose, providing extensive coverage and regular updates. However, manual analysis of these images is time-consuming and resource-intensive.

Semantic segmentation, a computer vision technique that assigns a class label to each pixel, offers a robust solution for identifying and classifying built-up areas. The advent of Deep Learning (DL), particularly Convolutional Neural Networks (CNNs), has enabled significant advancements in semantic segmentation by allowing models to automatically learn relevant image features directly from data.

This work aims to develop an automated system based on deep learning to segment built-up areas in satellite imagery. The study focuses on utilising a UNet-based segmentation model[1] and experimenting with various encoders to identify anthropised surfaces accurately. The performance of these encoders, including VGGNet, EfficientNet, ResNet, and Mix Vision Transformer (MiT), is evaluated using metrics such as Intersection over Union (IoU), F1-score, and balanced accuracy.

II. Methods

The core of the proposed system is a UNet architecture. The encoders tested include VGGNet, various ResNet versions (ResNet18, ResNet34, ResNet50), EfficientNet versions (B0-B2), and Mix Vision Transformer (MiT) versions (MiT-B0 to MiT-B4).

MapSAT, an Italian geospatial company, provided the dataset. It comprises satellite images from four Italian geographical areas: Grosseto, Gorizia, Caserta, and a portion of Lombardy. The ground truths were binary masks identifying two classes: built-up areas (positive class) and background/natural areas (negative class), as illustrated in Fig. **??**

Image preprocessing was a critical step. Due to the large size of the original satellite images (e.g., Grosseto: 8701x10960 pixels), they were partitioned into smaller 512x512 pixel tiles with a 50% overlap. The original 16-bit images were converted to 8-bit to reduce memory usage and processing time, which is beneficial for GPU VRAM limitations. A percentile-based normalisation (2nd and 98th percentiles) was applied to mitigate information loss from this conversion to enhance contrast and brightness.

The loss function employed was a combination of DiceLoss and FocalLoss[2], equally weighted. This choice helps address the significant class imbalance present in the dataset (87.49% negative class vs. 12.51% positive class). Hyperparameters

Fig. 1. Example input: a satellite image tile (left) and its ground truth segmentation mask for built-up areas (right, positive class in white).

included 70 epochs, a patience of 35 for early stopping, the Adam optimizer with a learning rate of 10^{-4}, and a batch size adjusted based on GPU capacity (typically 20).

III. Results

The experimental process was divided into two main phases.

This initial phase, *Phase 1*, utilized only spring images, with less accurate ground truths, notably lacking road segmentations. The dataset was split geographically: Grosseto and Lombardy for training, Caserta for validation, and Gorizia for testing. Despite the ground truth limitations, this phase aimed to assess the general feasibility of using neural networks. Among the tested encoders, reported in Tab I, EfficientNet-B0 demonstrated the best performance based on quantitative metrics like IoU (0.686) and F1-score (0.807).

TABLE I
Performance of different encoders in the *Phase 1*

Encoder	Balanced Acc.	IoU	Precision	Recall	F1-score
VGG13	91.8%	0.678	0.765	0.852	0.802
VGG19	92.6%	0.675	0.751	0.866	0.800
ResNet18	92.1%	0.683	0.762	0.856	0.804
ResNet34	**93.8%**	0.675	0.730	**0.891**	0.798
ResNet50	91.3%	0.670	0.760	0.840	0.795
EfficientNet-B0	91.3%	**0.686**	**0.777**	0.841	**0.807**
EfficientNet-B1	92.5%	0.669	0.739	0.867	0.796
EfficientNet-B2	89.5%	0.624	0.716	0.803	0.750
MiT-B0	92.5%	0.663	0.731	0.869	0.789
MiT-B1	92.7%	0.670	0.740	0.871	0.795

Following the promising results of *Phase 1* and the subsequent refinement of ground truths (including roads), this phase incorporated both spring and autumn images from the entire dataset. The geographical split was: Grosseto and Gorizia for training, Caserta for validation, and Lombardy for testing (with qualitative evaluation for Lombardy due to its unrefined annotations).

The results from *Phase 2*, in Tab II, showed a significant improvement in segmentation quality across all models, underscoring the impact of accurate ground truth data. The MiT-based encoders generally outperformed others. Specifically, UNet with the MiT-B2 encoder emerged as the top-performing model, achieving the best compromise between accuracy and segmentation precision.

TABLE II
Performance of different encoders in the *Phase 2*

Encoder	Balanced Acc.	IoU	Precision	Recall	F1-score
VGG16	95.7%	0.723	0.757	0.941	0.839
ResNet18	95.7%	0.737	0.774	0.939	0.848
ResNet34	95.8%	0.726	0.759	0.943	0.841
EfficientNet-B0	95.9%	0.732	0.764	0.945	0.845
EfficientNet-B1	96.3%	0.729	0.755	0.954	0.843
EfficientNet-B2	96.1%	0.722	0.750	0.951	0.838
MiT-B0	95.8%	0.745	0.782	0.939	0.854
MiT-B1	96.1%	0.735	0.766	0.948	0.847
MiT-B2	96.0%	**0.757**	**0.795**	0.941	**0.862**
MiT-B3	**96.5%**	0.740	0.767	0.956	0.851
MiT-B4	96.4%	0.729	0.753	**0.957**	0.843

A qualitative comparison between the best models from *Phase 1* (EfficientNet-B0) and *Phase 2* (MiT-B2) on Caserta images demonstrated the advancements, with MiT-B2 correctly segmenting areas, particularly roads, that were missed or poorly defined by the earlier model.

Furthermore, a qualitative assessment by MapSAT experts compared ResNet18, EfficientNet-B0, and MiT-B2 from *Phase 2*. MiT-B2 was noted for its good performance on built-up areas and roads, though it slightly overestimates built-up areas by including some green spaces. It showed behaviour comparable to ResNet18 for parking lots and cemeteries, and good performance against bare soil confusion. However, it performed less well in photovoltaic fields than the others. Compared to MapSAT's previous semi-automated and manually refined procedures, the deep learning approach, particularly with MiT-B2, markedly improved in reducing false alarms, especially the misclassification of bare soil and tilled fields as impervious surfaces. The new model also successfully identified roads that were often missed by the previous method.

IV. Discussion and Conclusion

This study successfully demonstrated the efficacy of a UNet-based deep learning approach for segmenting anthropized areas in satellite imagery. The experimental results highlighted that the MiT-B2 encoder provided the best accuracy and segmentation precision balance.

Compared to previous methodologies used by MapSAT, the developed deep learning system, without manual post-processing, significantly reduced false positives and more accurately delineated various types of anthropized surfaces, including road networks.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention, MICCAI 2015*. Springer, 2015, pp. 234–241.

[2] R. Azad, M. Heidary, K. Yilmaz, M. Hüttemann, S. Karimijafarbigloo, Y. Wu, A. Schmeink, and D. Merhof, "Loss functions in the era of semantic segmentation," *arXiv preprint arXiv:2312.05391*, 2023.